On Sequential Estimation and Prediction for Discrete Time Series

Gusztáv MORVAI and Benjamin WEISS

Stochastics and Dynamics, Vol. 7, No. 4. pp. 417-437, 2007

Abstract

The problem of extracting as much information as possible from a sequence of observations of a stationary stochastic process $X_0, X_1, ...X_n$ has been considered by many authors from different points of view. It has long been known through the work of D. Bailey that no universal estimator for $\mathbf{P}(X_{n+1}|X_0, X_1, ...X_n)$ can be found which converges to the true estimator almost surely. Despite this result, for restricted classes of processes, or for sequences of estimators along stopping times, universal estimators can be found. We present here a survey of some of the recent work that has been done along these lines.

1 Introduction

In a short communication that appeared in the Proceedings of the First International IEEE-USSR Information Workshop [7], Tom Cover formulated a number of problems that have generated a substantial literature during the past thirty years. We plan to survey a portion of these works, biased to be sure by our own intersets. We begin by quoting from Cover's paper and recalling his first two questions:

" 1. A Question on the Prediction of Ergodic Processes

The statement that "we can learn the statistics of an ergodic process from a sample function with probability 1" is being investigated for operational significance.

Let $\{X_n\}_{-\infty}^{\infty}$ be a stationary binary ergodic process with conditional probability distributions $p(x_{n+1}|x_n, \ldots, x_1)$, $n = 1, 2, \ldots$. We know that we can learn the statistics with probability 1, but can we learn p $\underline{fast}$ enough? In other words, does there exist an estimate $\hat{p} : X \times X^\star \to [0, 1]$, $X^\star =$ collection of all finite strings, for which

$$\hat{p}(X_{n+1}|X_n, \ldots, X_1) - p(X_{n+1}|X_n, \ldots, X_1) \to 0$$

with probability 1?

Does there also exist a predictor $\hat{p}$ yielding the convergence of

$$\hat{p}(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) \to p(X_0|X_{-1}, X_{-2}, \ldots)?$$

Since the statement of this problem, Bailey and Ornstein have obtained some as yet unpublished results on this question that indicate a negative answer to the first question and a positive answer to the second."

Since the processes are stationary, the (second) backward prediction problem is equivalent to the (first) forward prediction problem as far as convergence in probability is concerned. However, for almost sure results it turns out that they are far from being the same. Ornstein [30] gave a rather complicated algorithm for the backward prediction problem whereas Bailey

provided a proof for the nonexistence of a universal algorithm guaranteeing almost sure convergence in the forward estimation problem. To do this, Bailey in [5], assuming the existence of a universal algorithm, used the Ornstein's technique of cutting and stacking [31] for the construction of a "counterexample" process for which the algorithm fails to converge (see Shields [34] for more details on this method).

The problem came to life again in the late eighties with the work of Ryabko [33]. He used a simpler technique, namely - relabelling a countable state Markov chain, in order to prove the nonexistence of a universal estimator for Cover's first problem (cf. also Györfi, Morvai and Yakowitz [11]). In addition there was a growing interest in universal algorithms of various kinds in information theory and elsewhere, see Feder and Merhav [10] for a survey.

Three approaches evolved in an attempt to obtain positive results for the problem of forward estimation in the face of Bailey's theorem.

The first modifies the almost sure convergence to convergence in probability or almost sure convergence of the Cesaro averages. This was done already by Bailey in his thesis. Cf. Algoet [2, 3] and Weiss [36].

The second gives up on trying to estimate the distribution of the next output at all time moments n, and concentrates on guaranteeing prediction only at certain stopping times, while the third restricts the class of processes for which the scheme is shown to succeed.

Our interest in this circle of ideas began with the PhD thesis of the first author [15] in which he gave an algorithm for the backward prediction that was much simpler than Ornstein's original scheme (cf. Morvai, Yakowitz and Györfi [27]). Before describing briefly the contents of the survey we will present this scheme with a sketch of the proof of its validity. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from $\mathcal{X} = \{0, 1\}$. (Note that all stationary time series $\{X_n\}_{n=0}^{\infty}$ can be thought to be a two sided time series, that is, $\{X_n\}_{n=-\infty}^{\infty}$.) For notational convenience, let $X_m^n = (X_m, \ldots, X_n)$, where $m \leq n$.

Here is the algorithm. For $k = 1, 2, \ldots$, define sequences λ_{k-1} and τ_k recursively. Set $\lambda_0 = 1$ and let τ_k be the time between the occurrence of the pattern $X_{-\lambda_{k-1}}^{-1}$ at time -1 and the last occurrence of the same pattern prior to time -1. Formally, let

$$\tau_k = \min\{t > 0 : X_{-\lambda_{k-1}-t}^{-1-t} = X_{-\lambda_{k-1}}^{-1}\}.$$

Put

$$\lambda_k = \tau_k + \lambda_{k-1},$$

where λ_k is the length of the pattern

$$X_{-\lambda_k}^{-1} = X_{-\lambda_{k-1}-\tau_k}^{-1-\tau_k} X_{-\tau_k}^{-1}.$$

The observed vector $X_{-\lambda_{k-1}}^{-1}$ almost surely takes a value of positive probability; thus by stationarity, the string $X_{-\lambda_{k-1}}^{-1}$ must appear in the sequence $X_{-\infty}^{-2}$ almost surely. One denotes the kth estimate of $P(X_0 = 1|X_{-\infty}^{-1})$ by P_k, and defines it to be

$$P_k = \frac{1}{k} \sum_{j=1}^{k} X_{-\tau_j}.$$

As in Ornstein [30], the estimate P_k is calculated from observations of random size. Here the random sample size is λ_k. To obtain a fixed sample-size $0 < t < \infty$ version, we apply the same method as in Algoet [1], that is, let κ_t be the maximum of integers k for which $\lambda_k \leq t$. Formally,

$$\kappa_t = \max\{k \geq 0 : \lambda_k \leq t\}.$$

Now put

$$\hat{P}_{-t} = P_{\kappa_t}.$$

The following theorem was established in the PhD thesis of Morvai [15].

Theorem 1.1 *(Morvai [15]) For any stationary and ergodic binary time series* $\{X_n\}$,

$$\lim_{t\to\infty} \hat{P}_{-t} = P(X_0 = 1|X_{-\infty}^{-1}) \quad \textit{almost surely.}$$

Proof. We have

$$\begin{aligned} & P_k - P(X_0 = 1|X_{-\infty}^{-1}) \\ = & \frac{1}{k}\sum_{j=1}^{k}[X_{-\tau_j} - P(X_{-\tau_j} = 1|X_{-\lambda_{j-1}}^{-1})] \\ + & \frac{1}{k}\sum_{j=1}^{k} P(X_{-\tau_j} = 1|X_{-\lambda_{j-1}}^{-1}) - P(X_0 = 1|X_{-\infty}^{-1}). \end{aligned}$$

Observe that the first term is an average of a bounded martingale difference sequence and by Azuma's exponential bound for bounded martingale differences [4] we get that the first term tends to zero. Morvai showed in his PhD thesis that

$$P(X_{-\tau_j} = 1|X_{-\lambda_{j-1}}^{-1}) = P(X_0 = 1|X_{-\lambda_{j-1}}^{-1}).$$

This observation is the key to handling the second term:

$$\begin{aligned} & \frac{1}{k}\sum_{j=1}^{k} P(X_{-\tau_j} = 1|X_{-\lambda_{j-1}}^{-1}) - P(X_0 = 1|X_{-\infty}^{-1}) \\ = & \frac{1}{k}\sum_{j=1}^{k} P(X_0 = 1|X_{-\lambda_{j-1}}^{-1}) - P(X_0 = 1|X_{-\infty}^{-1}). \end{aligned}$$

By the martingale convergence theorem,

$$P(X_0 = 1|X_{-\lambda_{j-1}}^{-1}) \to P(X_0 = 1|X_{-\infty}^{-1}) \text{ almost surely,}$$

and since ordinary convergence implies Cesaro convergence this completes the proof of the theorem. □

In this survey we will restrict ourselves to finite or countably valued processes. Some of the directions that we survey have been generalized to real valued processes and some even to processes taking values in more general metric spaces. Some of the key papers in these directions are Algoet [1, 2, 3], Morvai et. al. [27, 26], Weiss [36] and Nobel [28].

We turn now to a brief description of the contents of our survey. In §2 we will describe some classes of processes that will play an important role for us. Next §3 will contain a scheme for forward prediction at all n which can be shown to converge to the optimal prediction for the class of processes with continuous conditional probabilities. This class includes of course k-step Markov chains for any k.

In §4 we turn to a description of a sequence of stopping times together with estimators which converge along that sequence to the conditional probability estimator for all processes. This sequence of stopping times grows rather quickly and we give a sequence with a slower growth rate but we can demonstrate the convergence only for processes whose conditional probabilities are almost surely continuous. Then in §5 for finitarily Markovian processes we give stopping times with an even slower growth rate. The following section considers this class in more detail with respect to the problem of estimating the length of the memory word that occurs as the context at time n.

We conclude with a series of constructions and examples in §§$7-9$ that show the optimality of many of these results. Along the way several open questions are mentioned since much remains to be done before we achieve a complete understanding of what is possible and what is not.

2 Preliminaries - Classes of Stochastic Processes

Let $\mathcal{X}$ be discrete (finite or countably infinite) alphabet. Let $\{X_n\}$ be a stationary and ergodic time series.
For notational convenience let $p(x_{-k}^0)$ and $p(y|x_{-k}^0)$ denote the distribution $P(X_{-k}^0 = x_{-k}^0)$ and the conditional distribution $P(X_1 = y|X_{-k}^0 = x_{-k}^0)$, respectively.

Definition 1. For a stationary time series $\{X_n\}$ the (random) length $K(X_{-\infty}^0)$ of the memory of the sample path $X_{-\infty}^0$ is the smallest possible $0 \leq K < \infty$ such that for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-K-i+1}^{-K} \in \mathcal{X}^i$

$$p(y|X_{-K+1}^0) = p(y|z_{-K-i+1}^{-K}, X_{-K+1}^0)$$

provided $p(z_{-K-i+1}^{-K}, X_{-K+1}^0, y) > 0$, and $K(X_{-\infty}^0) = \infty$ if there is no such K.

Note that we denote the random variables by capital letters and particular realizations by lower case letters. For example, $p(y|X_{-K+1}^0)$ is denoting the random variable which is a function of the random variables X_{-K+1}^0 taking the value $P(X_1 = y|X_{-k}^0 = x_{-k}^0)$ when $X_{-k}^0 = x_{-k}^0$.

Definition 2. The stationary time series $\{X_n\}$ is said to be finitarily Markovian if $K(X_{-\infty}^0)$ is finite (though not necessarily bounded) almost surely.

This class includes of course all finite order Markov chains but also many other processes such as the finitarily determined processes of Kalikow, Katznelson and Weiss [13], which serve to represent all isomorphism classes of zero entropy processes. For some concrete examples that are not Markovian consider the following example:

Example 1. Let $\{M_n\}$ be any stationary and ergodic first order Markov chain with finite or countably infinite state space S. Let $s \in S$ be an arbitrary state with $P(M_1 = s) > 0$. Now let $X_n = I_{\{M_n = s\}}$. By Shields ([35] Chapter I.2.c.1), the binary time series $\{X_n\}$ is stationary and ergodic. It is also finitarily Markovian. (Indeed, the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ does not depend on values beyond the first (going backwards) occurrence of one in $X_{-\infty}^0$ which identifies the first (going backwards) occurrence of state s in the Markov chain $\{M_n\}$.) The resulting time series $\{X_n\}$ is not a Markov chain of any order in general. (Indeed, consider the Markov chain $\{M_n\}$ with state space $S = \{0, 1, 2\}$ and transition probabilities $P(X_2 = 1|X_1 = 0) = P(X_2 = 2|X_1 = 1) = 1$, $P(X_2 = 0|X_1 = 2) = P(X_2 = 1|X_1 = 2) = 0.5$. This yields a stationary and ergodic Markov chain $\{M_n\}$, cf. (Example I.2.8

in Shields [35]. Clearly, the resulting time series $X_n = I_{\{M_n=0\}}$ will not be Markov of any order. The conditional probability $P(X_1 = 0|X_{-\infty}^0)$ depends on whether until the first (going backwards) occurrence of one you see even or odd number of zeros.) These examples include all stationary and ergodic binary renewal processes with finite expected inter-arrival times, a basic class for many applications. (A stationary and ergodic binary renewal process is defined as a stationary and ergodic binary process such that the times between occurrences of ones are independent and identically distributed with finite expectation, cf. Chapter I.2.c.1 in Shields [35]).

Let $\mathcal{X}^{*-}$ be the set of all one-sided sequences, that is,

$$\mathcal{X}^{*-} = \{(\ldots, x_{-1}, x_0) : x_i \in \mathcal{X} \text{ for all } -\infty < i \leq 0\}.$$

Let $f : \mathcal{X} \to (-\infty, \infty)$ be bounded, otherwise arbitrary. Define the function $F : \mathcal{X}^{*-} \to (-\infty, \infty)$ as

$$F(x_{-\infty}^0) = E(f(X_1)|X_{-\infty}^0 = x_{-\infty}^0).$$

E.g. if $f(x) = 1_{\{x=z\}}$ for a fixed $z \in \mathcal{X}$ then $F(y_{-\infty}^0) = P(X_1 = z|X_{-\infty}^0 = y_{-\infty}^0)$. If $\mathcal{X}$ is countably infinite subset of the reals and $f(x) = x$ then $F(y_{-\infty}^0) = E(X_1|X_{-\infty}^0 = y_{-\infty}^0)$.
Define the distance $d^*(\cdot, \cdot)$ on $\mathcal{X}^{*-}$ as follows. For $x_{-\infty}^0$, $y_{-\infty}^0 \in \mathcal{X}^{*-}$ let

$$d^*(x_{-\infty}^0, y_{-\infty}^0) = \sum_{i=0}^{\infty} 2^{-i-1} 1_{\{x_{-i} \neq y_{-i}\}}.$$

Definition 2.1 *We say that $F(X_{-\infty}^0)$ is continuous if a version of the function $F(X_{-\infty}^0)$ on the whole set $\mathcal{X}^{*-}$ is continuous with respect to metric $d^*(\cdot, \cdot)$.*

As we have already mentioned any k-step Markov chain satisfies this, but there are also many examples with unbounded memory. S. Kalikow showed

in [12] that the class can also be characterized as those processes which can be constructed as random Markov chains. In this procedure, given a past $X_{-\infty}^0$ one invokes an auxiliary independent process which chooses a random memory length K and then X_1 is chosen according to a fixed transition table from $\mathcal{X}^K$ to $\mathcal{X}$.

Definition 2.2 *We say that $F(X_{-\infty}^0)$ is almost surely continuous if for some set $C \subseteq \mathcal{X}^{*-}$ which has probability one a version of the function $F(X_{-\infty}^0)$ restricted to this set C is continuous with respect to metric $d^*(\cdot, \cdot)$.*

This class is strictly larger than the processes with continuous conditional distributions. It contains many of the examples that have been used to demonstrate the limitations of universal schemes. In particular, it contains the class of finitary Markov processes where the usual continuity may not hold (cf. Morvai and Weiss [17]).

3 Forward estimation for processes with continuous conditional distributions

For simplicity we will restrict our detailed presentation to the case where $\{X_n\}$ is a stationary and ergodic binary time series. As we have remarked, since we are interested primarily in pointwise results the restriction to ergodic processes doesn't lead to any loss of generality, while the extension to finite state processes is completely routine. Our goal is to estimate the conditional probability $P(X_{n+1} = 1|X_0^n)$ knowing only the samples X_0^n but not the nature of the process.

The following algorithm which was introduced in Morvai and Weiss [18] has several nice features. For processes with continuous conditional distribution the algorithm will almost surely give better and better prediction for X_{n+1} while for all other processes some type of convergence will obtain. For

$k \geq 1$ define the random variables $\tau_i^k(n)$ which indicate where the k-block X_{n-k+1}^n occurs previously in the time series $\{X_n\}$. Formally we set $\tau_0^k(n) = 0$ and for $i \geq 1$ let

$$\tau_i^k(n) = \min\{t > \tau_{i-1}^k(n) : X_{n-k+1-t}^{n-t} = X_{n-k+1}^n\}.$$

Let $K_n \geq 1$ and $J_n \geq 1$ be sequences of nondecreasing positive integers tending to ∞ which will be fixed later.
Define κ_n as the largest $1 \leq k \leq K_n$ such that there are at least J_n occurrences of the block X_{n-k+1}^n in the data segment X_0^n, that is,

$$\kappa_n = \max\{1 \leq k \leq \ K_n : \tau_{J_n}^k(n) \leq n - k + 1\}$$

if there is such k and 0 otherwise.
Define λ_n as the number of occurrences of the block $X_{n-\kappa_n+1}^n$ in the data segment X_0^n, that is,

$$\lambda_n = \max\{1 \leq j : \ \tau_j^{\kappa_n} \leq n - \kappa_n + 1\}$$

if $\kappa_n > 0$ and zero otherwise. Observe that if $\kappa_n > 0$ then $\lambda_n \geq J_n$.
Our estimate g_n for $P(X_{n+1} = 1|X_0^n)$ is defined as $g_0 = 0$ and for $n \geq 1$,

$$g_n = \frac{1}{\lambda_n} \sum_{i=1}^{\lambda_n} X_{n-\tau_i^{\kappa_n}(n)+1}$$

if $\kappa_n > 0$ and zero otherwise.

Theorem (Morvai and Weiss [18]) *Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite alphabet $\mathcal{X}$. Assume $K_n = \max(1, \lfloor 0.1 \log_{|\mathcal{X}|} n \rfloor)$ and $J_n = \max(1, \lceil n^{0.5} \rceil)$. Then*
(A) *if the conditional expectation $P(X_1 = 1|X_{-\infty}^0)$ is continuous with respect to metric $d^*(\cdot, \cdot)$ then*

$$\lim_{n\to\infty} |g_n - P(X_{n+1} = 1|X_0^n)| = 0 \quad \textit{almost surely,}$$

(B) *without any continuity assumption,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} |g_i - P(X_{i+1} = 1|X_0^i)| = 0 \quad \textit{almost surely,}$$

(C) *without any continuity assumption, for arbitrary* $\epsilon > 0$*,*

$$\lim_{n\to\infty} P(|g_n - P(X_{n+1} = 1|X_0^n)| > \epsilon) = 0.$$

Remarks:

We note that from the proof of Ryabko [33] and Györfi, Morvai, Yakowitz [11] it is clear that the continuity condition in the first part of the Theorem can not be relaxed. Even for the class of all stationary and ergodic binary time-series with merely almost surely continuous conditional probability $P(X_1 = 1|\ldots, X_{-1}, X_0)$ one can not achieve the convergence as in part (A).

We do not know if the shifted version of our proposed scheme g_n solves the backward estimation problem or not. That is, in the case when g_n is evaluated on $(X_{-n}, \ldots, X_0)$ rather than on $(X_0, \ldots, X_n)$, we expect convergence to be hold for all processes but we have been unable to prove this.

It is known that when the algorithms of Ornstein [30], Algoet [1], Morvai Yakowitz and Györfi [27] for the backward estimation problem are shifted forward parts (B) and (C) hold. For part (C) this is immediate from stationarity while for part (B) it follows from a generalized ergodic theorem, usually attributed to Breiman, but first proved by Maker [14]. Thus there is no novelty in the existence of some scheme with these properties. However, for the above algorithm all three properties hold. We should also point out that if one knows that the process is k-step Markov for some fixed k then of course it is not very hard to see that that the empirical distributions of the $k+1$-blocks converge almost surely by the ergodic theorem and this easily forms the basis of a scheme which will succeed in the forward prediction of these processes.

4 Estimating Along Stopping Times

The forward prediction problem for a binary time series $\{X_n\}_{n=0}^{\infty}$ is to estimate the probability that $X_{n+1} = 1$ based on the observations X_i, $0 \le i \le n$ without prior knowledge of the distribution of the process $\{X_n\}$. It is known that this is not possible if one estimates at all values of n. Morvai [16] presented a simple procedure which will attempt to make such a prediction infinitely often at carefully selected stopping times chosen by the algorithm. The growth rate of the stopping times can be determined. Here is his scheme.

Let $\{X_n\}_{n=-\infty}^{\infty}$ denote a two-sided stationary and ergodic binary time series. For $k = 1, 2, \ldots$, define the sequences $\{\tau_k\}$ and $\{\lambda_k\}$ recursively. Set $\lambda_0 = 0$. Let

$$\tau_k = \min\{t > 0 : X_t^{\lambda_{k-1}+t} = X_0^{\lambda_{k-1}}\}$$

and

$$\lambda_k = \tau_k + \lambda_{k-1}.$$

(By stationarity, the string $X_0^{\lambda_{k-1}}$ must appear in the sequence X_1^{∞} almost surely.) The kth estimate of $P(X_{\lambda_k+1} = 1|X_0^{\lambda_k})$ is denoted by P_k, and is defined as

$$P_k = \frac{1}{k-1}\sum_{j=1}^{k-1} X_{\lambda_j+1}.$$

Theorem 4.1 *(Morvai [16]) For all stationary and ergodic binary time series* $\{X_n\}$*,*

$$\lim_{k\to\infty}\Big(P_k - P(X_{\lambda_k+1} = 1|X_0^{\lambda_k})\Big) = 0 \quad \textit{almost surely.}$$

For some extensions of the algorithm see Morvai and Weiss [19].

One of the drawbacks of this scheme is that the growth of the stopping times $\{\lambda_k\}$ is rather rapid.

Theorem 4.2 *(Morvai [16]) Let $\{X_n\}$ be a stationary and ergodic binary time series. Suppose that $H > 0$ where*

$$H = \lim_{n\to\infty} -\frac{1}{n+1} E \log p(X_0, \ldots, X_n)$$

is the process entropy. Let $0 < \epsilon < H$ be arbitrary. Then for k large enough,

$$\lambda_k(\omega) \geq c^{c^{\cdot^{\cdot^{c}}}} \text{ almost surely,}$$

where the height of the tower is $k - d$, $d(\omega)$ is a finite number which depends on ω, and $c = 2^{H-\epsilon}$.

Morvai and Weiss [17] exhibited an estimator which is consistent on a certain stopping time sequence for a restricted class of stationary time series but which has a much slower rate of growth.

Define the stopping times now as follows. Set $\zeta_0 = 0$. For $k = 1, 2, \ldots$, define sequence η_k and ζ_k recursively. Let

$$\eta_k = \min\{t > 0 : X_{\zeta_{k-1}-(k-1)+t}^{\zeta_{k-1}+t} = X_{\zeta_{k-1}-(k-1)}^{\zeta_{k-1}}\} \text{ and } \zeta_k = \zeta_{k-1} + \eta_k.$$

One denotes the kth estimate of $P(X_{\zeta_k+1} = 1|X_0^{\zeta_k})$ by g_k, and defines it to be

$$g_k = \frac{1}{k} \sum_{j=0}^{k-1} X_{\zeta_j+1}.$$

Theorem 4.3 *(Morvai and Weiss [17]) Let $\{X_n\}$ be a stationary binary time series. Then*

$$\lim_{k\to\infty} \left| g_k - P(X_{\zeta_k+1} = 1|X_0^{\zeta_k}) \right| = 0 \quad \text{almost surely}$$

provided that the conditional probability $P(X_1 = 1|X_{-\infty}^0)$ is almost surely continuous.

Remark. We note that for all stationary binary time-series, the estimation scheme described above is consistent in probability.

Next we will give some universal estimates for the growth rate of the stopping times ζ_k in terms of the entropy rate of the process. This is natural since the ζ_k are defined by recurrence times for blocks of length k, and these are known to grow exponentially with the entropy rate.

Theorem 4.4 *(Morvai and Weiss [17]) Let $\{X_n\}$ be a stationary and ergodic binary time series. Then for arbitrary $\epsilon > 0$,*

$$\zeta_k < 2^{k(H+\epsilon)} \quad \textit{eventually almost surely,}$$

where H denotes the entropy rate associated with time series $\{X_n\}$.

This upper bound is much more favourable than the lower bound in Morvai [16]. For some extensions of this algorithm see Morvai and Weiss [24].

5 Some Improvements for Finitarily Markovian Processes

Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic (not necessarily finitarily Markovian) time series taking values from a discrete (finite or countably infinite) alphabet $\mathcal{X}$. Morvai and Weiss [23] provided the following algorithm which improves the performance of the previous one in case the process turns out to be finitarily Markovian.

For $k \geq 1$, let $1 \leq l_k \leq k$ be a nondecreasing unbounded sequence of integers, that is, $1 = l_1 \leq l_2 \ldots$ and $\lim_{k\to\infty} l_k = \infty$.

Define auxiliary stopping times (similarly to Morvai and Weiss [17]) as follows. Set $\zeta_0 = 0$. For $n = 1, 2, \ldots$, let

$$\zeta_n = \zeta_{n-1} + \min\{t > 0 : X_{\zeta_{n-1}-(l_n-1)+t}^{\zeta_{n-1}+t} = X_{\zeta_{n-1}-(l_n-1)}^{\zeta_{n-1}}\}.$$

Note that if $l_n = n$ then one gets $\zeta_n = \eta_n$ in Morvai and Weiss [17]. The point here is that l_n may grow slowly.

Among other things, using ζ_n and l_n we can define a very useful process $\{\tilde{X}_n\}_{n=-\infty}^{0}$ as a function of X_0^∞ as follows. Let $J(n) = \min\{j \geq 1 : \ l_{j+1} > n\}$ and define

$$\tilde{X}_{-i} = X_{\zeta_{J(i)}-i} \ \text{ for } i \geq 0.$$

In order to estimate $K(\tilde{X}_{-\infty}^0)$ we need to define some explicit statistics.

Define

$$\Delta_k(\tilde{X}_{-k+1}^0) =$$
$$\sup_{1\leq i} \sup_{\{z_{-k-i+1}^{-k}\in\mathcal{X}^i, x\in\mathcal{X}: p(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0, x)>0\}} \left| p(x|\tilde{X}_{-k+1}^0) - p(x|(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0)) \right|.$$

We will divide the data segment X_0^n into two parts: $X_0^{\lceil \frac{n}{2}\rceil -1}$ and $X_{\lceil \frac{n}{2}\rceil}^n$. Let $\mathcal{L}_{n,k}^{(1)}$ denote the set of strings with length $k+1$ which appear at all in $X_0^{\lceil \frac{n}{2}\rceil -1}$. That is,

$$\mathcal{L}_{n,k}^{(1)} = \{x_{-k}^0 \in \mathcal{X}^{k+1} : \exists k \leq t \leq \lceil \frac{n}{2}\rceil - 1 : X_{t-k}^t = x_{-k}^0\}.$$

For a fixed $0 < \gamma < 1$ let $\mathcal{L}_{n,k}^{(2)}$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in $X_{\lceil \frac{n}{2}\rceil}^n$. That is,

$$\mathcal{L}_{n,k}^{(2)} = \{x_{-k}^0 \in \mathcal{X}^{k+1} : \#\{\lceil \frac{n}{2}\rceil + k \leq t \leq n : X_{t-k}^t = x_{-k}^0\} > n^{1-\gamma}\}.$$

Let

$$\mathcal{L}_k^n = \mathcal{L}_{n,k}^{(1)} \bigcap \mathcal{L}_{n,k}^{(2)}.$$

We define the empirical version of Δ_k as follows:

$$\hat{\Delta}_k^n(\tilde{X}_{-k+1}^0) = \max_{1\leq i\leq n} \max_{(z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0, x)\in\mathcal{L}_{k+i}^n} 1_{\{\zeta_{J(k)}\leq \lceil \frac{n}{2}\rceil -1\}}$$
$$\left| \frac{\#\{\lceil \frac{n}{2}\rceil + k \leq t \leq n : X_{t-k}^t = (\tilde{X}_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2}\rceil + k - 1 \leq t \leq n-1 : X_{t-k+1}^t = \tilde{X}_{-k+1}^0\}} \right.$$
$$\left. - \ \frac{\#\{\lceil \frac{n}{2}\rceil + k + i \leq t \leq n : X_{t-k-i}^t = (z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0, x)\}}{\#\{\lceil \frac{n}{2}\rceil + k + i - 1 \leq t \leq n-1 : X_{t-k-i+1}^t = (z_{-k-i+1}^{-k}, \tilde{X}_{-k+1}^0)\}} \right|.$$

Note that the cut off $1_{\{\zeta_{J(k)} \leq \lceil \frac{n}{2} \rceil - 1\}}$ ensures that $\tilde{X}^0_{-k+1}$ is defined from $X_0^{\lceil \frac{n}{2} \rceil - 1}$.
Observe, that by ergodicity, for any fixed k,

$$\liminf_{n \to \infty} \hat{\Delta}^n_k \geq \Delta_k \quad \text{almost surely.}$$

We define an estimate χ_n for $K(\tilde{X}^0_{-\infty})$ from samples X_0^n as follows. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Set $\chi_0 = 0$, and for $n \geq 1$ let χ_n be the smallest $0 \leq k_n < n$ such that $\hat{\Delta}^n_{k_n} \leq n^{-\beta}$.
Observe that if $\zeta_j \leq \lceil \frac{n}{2} \rceil - 1 < \zeta_{j+1}$ then $\chi_n \leq l_{j+1}$.

Here the idea is that if $K(\tilde{X}^0_{-\infty}) < \infty$ then χ_n will be equal to $K(\tilde{X}^0_{-\infty})$ eventually and if $K(\tilde{X}^0_{-\infty}) = \infty$ then $\chi_n \to \infty$.

Now we define the sequence of stopping times λ_n along which we will be able to estimate. Set $\lambda_0 = \zeta_0$, and for $n \geq 1$ if $\zeta_j \leq \lambda_{n-1} < \zeta_{j+1}$ then put

$$\lambda_n = \min\{t > \lambda_{n-1} : X^t_{t-\chi_t+1} = X^{\zeta_j}_{\zeta_j - \chi_t + 1}\}$$

and

$$\kappa_n = \chi_{\lambda_n}.$$

Observe that if $\zeta_j \leq \lambda_{n-1} < \zeta_{j+1}$ then $\zeta_j \leq \lambda_{n-1} < \lambda_n \leq \zeta_{j+1}$. If $\chi_{\lambda_{n-1}+1} = 0$ then $\lambda_n = \lambda_{n-1} + 1$. Note that λ_n is a stopping time and κ_n is our estimate for $K(\tilde{X}^0_{-\infty})$ from samples $X_0^{\lambda_n}$.

Let $f : \mathcal{X} \to (-\infty, \infty)$ be bounded. One denotes the nth estimate of $E(f(X_{\lambda_n+1})|X_0^{\lambda_n})$ from samples $X_0^{\lambda_n}$ by f_n, and defines it to be

$$f_n = \frac{1}{n} \sum_{j=0}^{n-1} f(X_{\lambda_j+1}).$$

Fix positive real numbers $0 < \beta, \gamma < 1$ such that $2\beta + \gamma < 1$, fix a sequence l_n that $1 = l_1 \leq l_2, \ldots, l_n \to \infty$ and fix a bounded function $f(\cdot) : \mathcal{X} \to (-\infty, \infty)$

and with these numbers, sequence and function define ζ_n, χ_n, κ_n, λ_n and $F(\cdot)$ as described in the previous section. For the resulting f_n we have the following theorem:

Theorem 5.1 *(Morvai and Weiss [23]) Let $\{X_n\}$ be a stationary and ergodic time series taking values from a finite or countably infinite set $\mathcal{X}$. If the conditional expectation $F(X_{-\infty}^0)$ is almost surely continuous then almost surely,*

$$\lim_{n\to\infty} f_n = F(\tilde{X}_{-\infty}^0) \quad \textit{and} \quad \lim_{n\to\infty} \left| f_n - E(f(X_{\lambda_n+1})|X_0^{\lambda_n}) \right| = 0.$$

For arbitrary $\delta > 0$, $0 < \epsilon_2 < \epsilon_1$, let $l_n = \min\left(n, \max\left(1, \lfloor \frac{2+\delta}{\epsilon_1-\epsilon_2} \log_2 n \rfloor\right)\right)$. Then

$$\lambda_n < n^{\frac{2+\delta}{\epsilon_1-\epsilon_2}(H+\epsilon_1)}$$

eventually almost surely, and the upper bound is a polynomial whenever the stationary and ergodic time series $\{X_n\}$ has finite entropy rate H.
If the stationary and ergodic time series $\{X_n\}$ turns out to be finitarily Markovian then

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = \frac{1}{p(\tilde{X}^0_{-K(\tilde{X}^0_{-\infty})+1})} < \infty \quad \textit{almost surely.}$$

Moreover, if the stationary and ergodic time series $\{X_n\}$ turns out to be independent and identically distributed then $\lambda_n = \lambda_{n-1} + 1$ eventually almost surely.

6 Estimation for Finitarily Markovian Processes

In this section we broaden the scope of the estimation question that we will discuss and describe first how well can we detect the presence of a memory

word in a finitarily Markovian process (cf. Morvai and Weiss [25]). This problem has been discussed often in the context of modelling processes. Here we will show how it relates to prediction questions.

Recall that K was the minimal length of the context that defines the conditional probability. We take up the problem of estimating the value of K, both in the backward sense and in the forward sense, where one observes successive values of $\{X_n\}$ for $n \geq 0$ and asks for the least value K such that the conditional distribution of X_{n+1} given $\{X_i\}_{i=n-K+1}^{n}$ is the same as the conditional distribution of X_{n+1} given $\{X_i\}_{i=-\infty}^{n}$. We will consider both finite and countably infinite alphabet size.

For the case of finite alphabet finite order Markov chains similar questions have been studied by Bühlman and Wyner in [6]. However, the fact that we want to treat countable alphabets complicates matters significantly. The point is that while finite alphabet Markov chains have exponential rates of convergence of empirical distributions, for countable alphabet Markov chains no universal rates are available at all.

This problem appears in Morvai and Weiss [21] where a universal estimator for the order of a Markov chain on a countable state space is given, and some of the techniques that are used in the proofs of the results described here have their origin in that paper. We note in passing, that in Morvai and Weiss [20] it is shown that there is no classification rule for discriminating the class of finitarily Markovian processes from other ergodic processes.

The key notion is that of a **memory word** which can be defined as follows.

Definition 6.1 *We say that w_{-k+1}^{0} is a memory word if for all $i \geq 1$, all $y \in \mathcal{X}$, all $z_{-k-i+1}^{-k} \in \mathcal{X}^i$*

$$p(y|w_{-k+1}^{0}) = p(y|z_{-k-i+1}^{-k}, w_{-k+1}^{0})$$

provided $p(z_{-k-i+1}^{-k}, w_{-k+1}^{0}, y) > 0$.

Define the set $\mathcal{W}_k$ of those memory words w_{-k+1}^0 with length k, that is,

$$\mathcal{W}_k = \{w_{-k+1}^0 \in \mathcal{X}^k : \ w_{-k+1}^0 \text{ is a memory word}\}.$$

Our first result is a solution of the backward estimation problem, namely determining the value of $K(X_{-\infty}^0)$ from observations of increasing length of the data segments X_{-n}^0. We will give in the next subsection a universal consistent estimator which will converge almost surely to the memory length $K(X_{-\infty}^0)$ for any ergodic finitarily Markovian process on a countable state space. The detailed proofs in Morvai and Weiss [25] are pretty explicit and given some information on the average length of a memory word and the extent to which the stationary distribution diffuses over the state space one could extract rates for the convergence of the estimators. We concentrate however, on the more universal aspects of the problem.

As is usual in these kinds of questions , the problem of forward estimation, namely trying to determine $K(X_{-\infty}^n)$ from successive observations of X_0^n is more difficult. The stationarity means that results in probability can be carried over automatically. However, almost sure results present serious problems as we have already said. For some more results in this circle of ideas of what can be learned about processes by forward observations see Ornstein and Weiss [32], Dembo and Peres [9], Nobel [29], and Csiszár and Talata [8].

Recently in Csiszár and Talata [8] the authors define a finite context to be a memory word w of minimal length, that is, no proper suffix of w is a memory word. An infinite context for a process is an infinite string with all finite suffix having positive probability but none of them being a memory word. They treat there the problem of estimating the entire context tree in case the size of the alphabet is finite. For a bounded depth context tree, the process is Markovian, while for an unbounded depth context tree the universal pointwise consistency result there is obtained only for the truncated trees which are again finite in size. This is in contrast to our results which deal with infinite alphabet size and consistency in estimating memory words

of arbitrary length. This is what forces us to consider estimating at specially chosen times.

In the second subsection we will present a scheme which depend upon a positive parameter ϵ, and we guarantee that density of times along which the estimates are being given have density at least $1-\epsilon$. The last two subsections are devoted to seeing how this memory length estimation can be applied to estimating conditional probabilities. We do this first for finitarily Markovian processes along a sequence of stopping times which achieve density $1-\epsilon$. We do not know if the ϵ can be dropped in this case for the estimation of conditional probabilities.

We can dispense with ϵ in the Markovian case. For this we use an earlier result of ours on a universal estimator for the order of a finite order Markov chain on a countable alphabet in order to estimate the conditional probabilities along a sequence of stopping times of density one.

6.1 Backward Estimation of the Memory Length for Finitarily Markovian Processes

Let $\{X_n\}$ be stationary and ergodic finitarily Markovian with finite or countably infinite alphabet.

In order to estimate $K(X_{-\infty}^0)$ we need to define some explicit statistics. The first is a measurement of the failure of w_{-k+1}^0 to be a memory word.

Define

$$\Delta_k(w_{-k+1}^0) = \sup_{1\le i} \sup_{\{z_{-k-i+1}^{-k}\in\mathcal{X}^i, x\in\mathcal{X}: p(z_{-k-i+1}^{-k}, w_{-k+1}^0, x)>0\}} \left| p(x|w_{-k+1}^0) - p(x|z_{-k-i+1}^{-k}, w_{-k+1}^0) \right|.$$

Clearly this will vanish precisely when w_{-k+1}^0 is a memory word. We need to define an empirical version of this based on the observation of a finite data segment X_{-n}^0. To this end first define the empirical version of the conditional

probability as

$$\hat{p}_n(x|w^0_{-k+1}) = \frac{\#\{-n+k-1 \le t \le -1 : X^{t+1}_{t-k+1} = (w^0_{-k+1}, x)\}}{\#\{-n+k-1 \le t \le -1 : X^t_{t-k+1} = w^0_{-k+1}\}}.$$

These empirical distributions, as well as the sets we are about to introduce are functions of X^0_{-n}, but we suppress the dependence to keep the notation manageable.

For a fixed $0 < \gamma < 1$ let $\mathcal{L}^n_k$ denote the set of strings with length $k+1$ which appear more than $n^{1-\gamma}$ times in X^0_{-n}. That is,

$$\mathcal{L}^n_k = \{x^0_{-k} \in \mathcal{X}^{k+1} : \#\{-n+k \le t \le 0 : X^t_{t-k} = x^0_{-k}\} > n^{1-\gamma}\}.$$

Finally, define the empirical version of Δ_k as follows:

$$\hat{\Delta}^n_k(w^0_{-k+1}) = \max_{1 \le i \le n} \max_{(z^{-k}_{-k-i+1}, w^0_{-k+1}, x) \in \mathcal{L}^n_{k+i}} \left| \hat{p}_n(x|w^0_{-k+1}) - \hat{p}_n(x|z^{-k}_{-k-i+1}, w^0_{-k+1}) \right|$$

Let us agree by convention that if the smallest of the sets over which we are maximizing is empty then $\hat{\Delta}^n_k = 0$. Observe, that by ergodicity, the ergodic theorem implies that almost surely the empirical distributions $\hat{p}$ converge to the true distributions p and so for any $w^0_{-k+1} \in \mathcal{X}^k$,

$$\liminf_{n \to \infty} \hat{\Delta}^n_k(w^0_{-k+1}) \ge \Delta_k(w^0_{-k+1}) \quad \text{almost surely.}$$

With this in hand we can give a test for w^0_{-k+1} to be a memory word. Let $0 < \beta < \frac{1-\gamma}{2}$ be arbitrary. Let $NTEST_n(w^0_{-k+1}) = YES$ if $\hat{\Delta}^n_k(w^0_{-k+1}) \le n^{-\beta}$ and NO otherwise. Note that $NTEST_n$ depends on X^0_{-n}.

Theorem 6.1 *(Morvai and Weiss [25]) Eventually almost surely, $NTEST_n(w^0_{-k+1}) = YES$ if and only if w^0_{-k+1} is a memory word.*

We define an estimate χ_n for $K(X^0_{-\infty})$ from samples X^0_{-n} as follows. Set $\chi_0 = 0$, and for $n \ge 1$ let χ_n be the smallest $0 \le k < n$ such that $NTEST_n(X^0_{-k+1}) = YES$ if there is such and n otherwise.

Theorem 6.2 *(Morvai and Weiss [25]) $\chi_n = K(X^0_{-\infty})$ eventually almost surely.*

6.2 Forward Estimation of the Memory Length for Finitarily Markovian Processes

Let $\{X_n\}$ be stationary and ergodic finitarily Markovian with finite or countably infinite alphabet.
Define $PTEST_n(w_{-k+1}^0)(X_0^n) = NTEST_n(w_{-k+1}^0)(T^n X_0^n)$ where T is the left shift operator.

Theorem 6.3 *(Morvai and Weiss [25]) Eventually almost surely, $PTEST_n(w_{-k+1}^0) = YES$ if and only if w_{-k+1}^0 is a memory word.*

Define a list of words $\{w(0), w(1), w(2), \ldots, w(n), \ldots\}$ such that all words of all lengths are listed and a word can not precede its suffix. Note that $w(0)$ is the empty word.
Now define sets of indices A_n^i as follows. Let $A_n^0 = \{0, 1, \ldots, n\}$ and for $i > 0$ define

$$A_n^i = \{|w(i)| - 1 \leq j \leq n : X_{j-|w(i)|+1}^j = w(i)\}. \tag{1}$$

Let $\epsilon > 0$ be fixed. Define $\theta_n(\epsilon) < n$ to be the minimal j such that

$$\frac{\left|\bigcup_{i \leq j : PTEST_n(w(i))=YES} A_n^i\right|}{n+1} \geq 1 - \epsilon/2 \tag{2}$$

and n otherwise. We estimate for the length of the memory of $X_{-\infty}^n$ looking backwards if $n \in \bigcup_{i \leq \theta_n(\epsilon), PTEST_n(w(i))=YES} A_n^i$. The set of n's for which this holds will be the set for which we estimate the memory and we denote this set by $\mathcal{N}$. Note that the event $n \in \mathcal{N}$ depends only on X_0^n, and thus $\mathcal{N}$ can be thought of as a sequence of stopping times.
We define for $n \in \mathcal{N}$,

$$\kappa_n = \min\{i \geq 0 : X_{n-|w(i)|+1}^n = w(i), PTEST_n(w(i)) = YES\}.$$

For $n \in \mathcal{N}$ define

$$\rho_n(X_0^n) = |w(\kappa_n)|.$$

Note that ρ_n, θ_n, κ_n and $\mathcal{N}$ depend on ϵ, however, we will not denote this dependence on epsilon explicitly.

Theorem 6.4 *(Morvai and Weiss [25]) Let $\epsilon > 0$ be fixed. Then for $n \in \mathcal{N}$,*

$$\rho_n = K(X_{-\infty}^n) \text{ eventually almost surely,} \tag{3}$$

and

$$\liminf_{n\to\infty} \frac{|\mathcal{N} \bigcap \{0, 1, \ldots, n-1\}|}{n} \geq 1 - \epsilon. \tag{4}$$

For $n \in \mathcal{N}$, $X_{n-\rho_n+1}^n$ appears at least $n^{-\gamma}$ times eventually almost surely.

6.3 Forward Estimation of the Conditional Probability for Finitarily Markovian Processes

Let $\{X_n\}$ be stationary and ergodic finitarily Markovian with finite or countably infinite alphabet. Now our goal is to estimate the conditional probability $P(X_{n+1} = x|X_0^n)$ on stopping times in a pointwise sense.

Let $\mathcal{N}$ be a sequence of stopping times such that eventually almost surely $X_{n-K(X_{-\infty}^n)+1}^n$ appears at least $n^{1-\gamma}$ times in X_0^n.

Let ρ_n be any estimate of the length of the memory from samples X_0^n such that $\rho_n - K(X_{-\infty}^n) \to 0$ on $\mathcal{N}$.

Define our estimate $\hat{q}_n(x)$ of the conditional probability $P(X_{n+1} = x|X_0^n)$ on $\mathcal{N}$ as

$$\hat{q}_n(x) = \frac{\#\{\rho_n - 1 \leq i < n : X_{i-\rho_n+1}^i = X_{n-\rho_n+1}^n, X_{n+1} = x\}}{\#\{\rho_n - 1 \leq i < n : X_{i-\rho_n+1}^i = X_{n-\rho_n+1}^n\}}.$$

Theorem 6.5 *(Morvai and Weiss [25]) On $n \in \mathcal{N}$,*

$$|\hat{q}_n(x) - P(X_{n+1} = x|X_0^n)| \to 0 \text{ almost surely.}$$

Corollary 6.1 *For the stopping times $\mathcal{N}$ and estimator ρ_n in Theorem 6.4, Theorem 6.5 holds and the density of $\mathcal{N}$ is at least $1 - \epsilon$.*

6.4 Forward Estimation of the Conditional Probability for Markov Processes

Let $\{X_n\}$ be a stationary and ergodic finite or countably infinite alphabet Markov chain with order K. Let $ORDEST_n$ be an estimator of the order from samples X_0^n such that $ORDEST_n \to K$ almost surely. Such an estimator can be found e.g. in Morvai and Weiss [21]. Let $n \in \mathcal{N}$ if $X_{n-ORDEST_n+1}^n$ appears at least $n^{1-\gamma}$ times in X_0^n. $\mathcal{N}$ is a sequence of stopping times. Let

$$\hat{q}_n(x) = \frac{\#\{ORDEST_n - 1 \leq i < n : X_{i-ORDEST_n+1}^i = X_{n-ORDEST_n+1}^n, X_{n+1} = x\}}{\#\{ORDEST_n - 1 \leq i < n : X_{i-ORDEST_n+1}^i = X_{n-ORDEST_n+1}^n\}}.$$

Theorem 6.6 *(Morvai and Weiss [25]) Assume $ORDEST_n$ equals the order eventually almost surely. Then on $n \in \mathcal{N}$,*

$$|\hat{q}_n(x) - P(X_{n+1} = x|X_{n-K}^n)| \to 0 \text{ almost surely.}$$

and

$$\liminf_{n\to\infty} \frac{|\mathcal{N} \bigcap \{0, 1, \ldots, n-1\}|}{n} = 1.$$

If the Markov chain turns out to take values from a finite set, then $\mathcal{N}$ takes as values all but finitely many positive integers.

7 Examples Illustrating Limitations

For the class of all stationary and ergodic binary Markov-chains of some finite order the forward estimation problem can be solved. Indeed, if the time series is a Markov-chain of some finite order, we can estimate the order and count frequencies of blocks with length equal to the order. Bailey showed that one can't test for being in the class, cf. Morvai and Weiss [20] also.

It is conceivable that one can improve the result of Morvai [16] or Morvai and Weiss [17] so that if the process happens to be Markovian then one eventually estimates at all times. It has been shown in Morvai and Weiss

[22] that this is not possible. This puts some new restrictions on what can be achieved in estimating along stopping times.

Theorem 7.1 *(Morvai and Weiss [22]) For any strictly increasing sequence of stopping times* $\{\lambda_n\}$ *such that for all stationary and ergodic binary Markov-chains with arbitrary finite order, eventually* $\lambda_{n+1} = \lambda_n + 1$*, and for any sequence of estimators* $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ *there is a stationary and ergodic binary time series* $\{X_n\}$ *with almost surely continuous conditional probability* $P(X_1 = 1|\ldots, X_{-1}, X_0)$*, such that*

$$P\left(\limsup_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - P(X_{\lambda_n+1} = 1|X_0, \ldots, X_{\lambda_n})| > 0\right) > 0.$$

Remark: Bailey [5] among other things proved that there is no sequence of functions $\{e_n(X_0^{n-1})\}$ which for all stationary and ergodic time series, if it turns out to be a Markov-chain, would be eventually 1 and 0 otherwise. (That is, there is no test for the Markov property.) This result does not imply ours. On the other hand, our result implies Bailey's. (Indeed, if there were a test for Markov-chains in the above sense, we could apply the estimator in Morvai [16] or Morvai and Weiss [17] if the time series is not a Markov-chain of some finite order, and if the time series is a Markov-chain of some finite order we can estimate the order of the Markov chain and count frequencies of blocks with length equal to the order.

Bailey [5] and Ryabko [33] proved less than our theorem. They proved the nonexistence of the desired estimator when the estimator should work for all stationary and ergodic binary time series and when all $\lambda_n = n$, that is, when we always require good prediction.

8 Memory Estimation for Markov Processes

In this section we shall examine how well can one estimate the local memory length for finite order Markov chains. In the case of finite alphabets this can

be done with stopping times that eventually cover all time epochs. (Indeed, assume $\{X_n\}$ is a Markov chain taking values from a finite set. Assume $ORDEST_n$ estimates the order in a pointwise sense from data X_0^n. Then let

$$\rho_n = \min\{0 \le t \le ORDEST_n : \ PTEST_n(X_{n-t+1}^n) = YES\}$$

if there is such t and 0 otherwise. Since $ORDEST_n$ eventually gives the right order and there are finitelly many possible strings with length not greater than the order thus $\rho_n = K(X_{-\infty}^n)$ eventually almost surely by Theorem 6.3.)

However, as soon as one goes to a countable alphabet, even if the order is known to be two and we are just trying to decide whether the X_n alone is a memory word or not, there is no sequence of stopping times which is guaranteed to succeed eventually and whose density is one, cf. Morvai and Weiss [25]. This shows that the ϵ in the preceding sections cannot be eliminated.

Theorem 8.1 *(Morvai and Weiss [25]) There are no strictly increasing sequence of stopping times $\{\lambda_n\}$ and estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$ taking the values one and two, such that for all countable alphabet Markov chains of order two:*

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$$

and

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - K(X_0^{\lambda_n})| = 0 \quad \textit{with probability one.}$$

9 Limitations for Binary Finitarily Markovian Processes

In the preceding section we showed that we cannot achieve density one in the forward memory length estimation problem even in the class of Markov chains on a countable alphabet. In this section we shall show something

similar in the class of binary (i.e. $0, 1$) valued finitarily Markov processes. We will assume that there is given a sequence of estimators and stopping times, (h_n, λ_n) that do succeed to estimate successfully the memory length for binary Markov chains of finite order and construct a finitarily Markovian binary process on which the scheme fails infinitely often. Here is a precise statement:

Theorem 9.1 *(Morvai and Weiss [25]) For any strictly increasing sequence of stopping times $\{\lambda_n\}$ and sequence of estimators $\{h_n(X_0, \ldots, X_{\lambda_n})\}$, such that for all stationary and ergodic binary Markov chains with arbitrary finite order, $\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$, and*

$$\lim_{n\to\infty} |h_n(X_0, \ldots, X_{\lambda_n}) - K(X_0^{\lambda_n})| = 0 \quad \textit{almost surely}$$

there is a stationary, ergodic finitarily Markovian binary time series such that on a set of positive measure of process realizations

$$h_n(X_0, \ldots, X_{\lambda_n}) \neq K(X_{-\infty}^{\lambda_n})$$

infinitely often.

In the final process X_n that we constructed in Morvai and Weiss [25] we have $P(K(X_{-\infty}^0) = k$ decays to zero exponentially fast and in particular is summable. It follows that with probability one eventually $K(X_0^n) \leq n$ so that the reason for our failure to estimate the order correctly is not coming about because we don't even see the memory word.

It is also worth pointing out the density of moments on which the estimator is failing is of density zero. It follows fairly easily from the ergodic theorem that if one is willing to tolerate such failures then a straightforward application of any backward estimation scheme will converge outside a set of density zero.

References

[1] P. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *Annals of Probability*, vol. 20, pp. 901–941, 1992. Correction: *ibid.* vol. 23, pp. 474–478, 1995.

[2] P. Algoet, "The strong low of large numbers for sequential decisions under uncertainity," *IEEE Transactions on Information Theory*, vol. 40, pp. 609–634, 1994.

[3] P. Algoet, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Transactions on Information Theory*, vol. 45, pp. 1165–1185, 1999.

[4] K. Azuma, "Weighted sums of certain dependent random variables," in *Tohoku Mathematical Journal,* vol. 37, pp. 357–367, 1967.

[5] D. H. Bailey, *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph. D. thesis, Stanford University, 1976.

[6] P. Bühlmann and A. J. Wyner, "Variable- length Markov chains," *Annals of Statistics* , vol. 27, pp. 480–513, 1999.

[7] T. M. Cover, "Open problems in information theory," in *1975 IEEE Joint Workshop on Information Theory*, pp. 35–36. New York: IEEE Press, 1975.

[8] I. Csiszár and Zs. Talata, "Context tree estimation for not necessarily finite memory processes via BIC and MDL," To appear in *IEEE Transactions on Information Theory,*.

[9] A. Dembo and Y. Peres, *A topological criterion for hypothesis testing* Annals of Stat. **22** (1994) 106-117.

[10] M. Feder and N. Merhav, "Universal prediction" *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, 1998.

[11] L. Györfi, G. Morvai, and S. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on Information Theory*, vol. 44, pp. 886–892, 1998.

[12] S. Kalikow "Random Markov processes and uniform martingales ," *Israel Journal of Mathematics*, vol. 71, pp. 33–54, 1990.

[13] S. Kalikow, Y. Katznelson and B. Weiss. "Finitarily deterministic generators for zero entropy systems", *Israel Journal of Mathematics*, vol. 79, pp. 33-45, 1992.

[14] Ph.T. Maker, "The ergodic theorem for a sequence of functions," *Duke Math. J.*, vol. 6, pp. 27–30, 1940.

[15] G. Morvai "Estimation of Conditional Distribution for Stationary Time Series " PhD Thesis, Technical University of Budapest, 1994.

[16] G. Morvai "Guessing the output of a stationary binary time series" In: Foundations of Statistical Inference, (Eds. Y. Haitovsky, H.R.Lerche, Y. Ritov), Physika-Verlag, pp. 207-215, 2003.

[17] G. Morvai and B. Weiss, "Forecasting for stationary binary time series" *Acta Applicandae Mathematicae*, vol. 79, 25–34, 2003.

[18] G. Morvai and B. Weiss, "Forward estimation for ergodic time series" *Ann. I.H.Poincaré Probabilités et Statistiqoes*, vol. 41, 859–870, 2005.

[19] G. Morvai and B. Weiss, "Inferring the conditional mean" *Theory of Stochastic Processes*, vol. 11, 112–120, 2005.

[20] G. Morvai and B. Weiss, "On classifying processes" *Bernoulli*, vol. 11, 523–532, 2005.

[21] G. Morvai and B. Weiss, "Order estimation of Markov chains," *IEEE Transactions on Information Theory*, vol. 51, pp. 1496-1497, 2005.

[22] G. Morvai and B. Weiss, "Limitations on intermittent forecasting" *Statistics and Probability Letters*, vol. 72, 285–290, 2005.

[23] G. Morvai and B. Weiss, "Prediction for discrete time series" *Probability Theory and Related Fields*, vol. 132, 1–12, 2005.

[24] G. Morvai and B. Weiss, "Intermittent estimation of stationary time series" *Test*, vol. 13, 525–542, 2004.

[25] G. Morvai and B. Weiss, "Estimating the memory for finitarily Markovian processes" *Ann. I.H.Poincaré Probabilités et Statistiqoes*, vol. 43, pp. 15-30, 2007.

[26] G. Morvai, S. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Transactions on Information Theory*, vol. 43, pp. 483-498, 1997.

[27] G. Morvai, S. Yakowitz, and L. Györfi, "Nonparametric inferences for ergodic, stationary time series," *Annals of Statistics.*, vol. 24, pp. 370–379, 1996.

[28] A. Nobel, " On optimal sequential prediction for general processes," *IEEE Trans. Inform. Theory*, vol. 49, no. 1, pp. 83–98, 2003.

[29] A. Nobel, "Limits to classification and regression estimation from ergodic processes," *Annals of Statistics*, vol. 27 pp. 262-273, 1999.

[30] D. S. Ornstein, "Guessing the next output of a stationary process," *Israel Journal of Mathematics,* vol. 30, pp. 292–296, 1978.

[31] D. S. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems.* Yale University Press, 1974.

[32] D.S. Ornstein and B. Weiss, "How sampling reveals a process," *The Annals of Probability*, vol. 18, pp. 905-930, 1990.

[33] B. Ya. Ryabko, "Prediction of random sequences and universal coding," *Problems of Inform. Trans.,* vol. 24, pp. 87-96, Apr.-June 1988.

[34] P.C. Shields, "Cutting and stacking: a method for constructing stationary processes," *IEEE Transactions on Information Theory,* vol. 37, pp. 1605–1614, 1991.

[35] P.C. Shields, *The Ergodic Theory of Discrete Sample Paths,* volume 13 of Graduate Studies in Mathematics. American Mathematical Society, Providence, 1996.

[36] B. Weiss, *Single Orbit Dynamics*, American Mathematical Society, 2000.